Journal of Dental Sciences

Volume 21 | Issue 2 Article 25

2026

Performance of large language models on image-based oral pathology questions from the Japanese National Dental Examination

Hikaru Watanabe

Osamu Uehara

Tetsuro Morikawa

Takayuki Kojima

Yoshihiro Abiko

Journal of Dental Sciences 21 (2026) 879–887

Available online at https://jds.ads.org.tw/journal/

Digital Commons

journal homepage: https://jds.ads.org.tw/journal/

Original Article

Performance of large language models on image-based oral pathology questions from the Japanese National Dental Examination

Hikaru Watanabe [a], Osamu Uehara [b], Tetsuro Morikawa [c], Takayuki Kojima [a], Takayuki Suga [d], Akira Toyofuku [d], Satoshi Takada [a], Yoshihiro Abiko [c*]

[a] *Department of Oral and Maxillofacial Surgery, Ohu University School of Dentistry, Fukushima, Japan*
[b] *Division of Disease Control and Molecular Epidemiology, Department of Oral Growth and Development, School of Dentistry, Health Sciences University of Hokkaido, Hokkaido, Japan*
[c] *Division of Oral Medicine and Pathology, Department of Human Biology and Pathophysiology, School of Dentistry, Health Sciences University of Hokkaido, Hokkaido, Japan*
[d] *Department of Psychosomatic Dentistry, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan*

Received 15 August 2025; Final revision received 22 August 2025
Available online 1 April 2026

KEYWORDS
Artificial intelligence;
Chat generative Pre-trained Transformer (ChatGPT);
Large language models;
Japanese national dental examination

Abstract *Background: /purpose*: Large language models (LLMs), such as Chat Generative Pre-trained Transformer (ChatGPT) and Gemini, have demonstrated promising capabilities for medical question-answering tasks. However, the diagnostic performance of LLMs in image-based oral pathologies remains largely unexplored. This study aimed to evaluate these capabilities using histopathological images obtained from the Japanese National Dental Examination.

Materials and methods: This study aimed to evaluate and compare the diagnostic accuracy and agreement of three LLMs (ChatGPT-4o [ChatGPT], Gemini 1.5 Pro [Gemini], and Claude 3.5 Sonnet [Claude]) on pathology image-based questions from the Japanese National Dental Examination.

Results: Gemini achieved the highest accuracy (61.4 %), followed by Claude (52.3 %), and ChatGPT (45.4 %). Gemini and ChatGPT exhibited significant differences ($P = 0.00054$). Cohen's kappa values indicated moderate agreement for all models, with Gemini showing the highest agreement ($\kappa = 0.599$). Accuracy varied across disease categories: Gemini excelled in squamous cell carcinoma (92.0 %) and salivary gland tumors, whereas Claude performed best on soft tissue lesions. The confusion matrix analysis revealed distinct misclassification patterns

* Corresponding author. Division of Oral Medicine and Pathology, Department of Human Biology and Pathophysiology, School of Dentistry, Health Sciences University of Hokkaido, 1757 Kanazawa, Ishikari-Tobetsu, Hokkaido, 061-0293, Japan.
E-mail address: yoshi-ab@hoku-iryo-u.ac.jp (Y. Abiko).

https://doi.org/10.1016/j.jds.2025.08.037

in each model, particularly between odontogenic tumors and cystic lesions.
Conclusion: LLMs demonstrated moderate diagnostic performance for image-based dental pathology questions, with Gemini demonstrating superior accuracy and consistency. Although promising decision support tools in education and clinical settings, LLMs still exhibit domain-specific limitations and require careful oversight. The integration of explainable artificial intelligence and real-world clinical validation is recommended for its safe and effective use.

Introduction

Artificial intelligence (AI) is receiving increasing attention across various fields of medicine, including pathological diagnosis.[1] The development of large language models (LLMs), such as Chat Generative Pre-trained Transformer (ChatGPT), has demonstrated their ability to generate human-like responses and support clinical decision making. With the recent integration of image analysis capabilities into LLMs, interest in their application to pathology has continued to increase.[2,3] Several studies have already examined the usefulness of LLMs in pathological diagnosis.[4] Notably, AI-based diagnostic tools have been employed in dermatology since the late 1980s.[5] More recently, studies comparing the diagnostic performance of dermatologists and deep learning models for skin cancer have shown that AI can achieve accuracy comparable to that of human specialists.[6]

Although the oral mucosa shares structural and pathological similarities with the skin, a limited number of studies have investigated the use of AI in diagnosing oral mucosal diseases.[7–10] Recently, only one study assessed the diagnostic performance of ChatGPT-4o in the field of oral pathology, particularly beyond oral mucosal diseases. To our knowledge, no comparative studies have evaluated the diagnostic accuracy of AI for oral pathology among different LLMs to date.[11] When using LLMs for pathological diagnosis, the specific part of the tissue specimen selected for image analysis can significantly affect the outcome. Additionally, images captured by expert pathologists are more likely to lead to accurate diagnoses than those captured by non-specialists. To properly evaluate the diagnostic accuracy, it is essential to use images with both high-quality and appropriately selected diagnostic regions. In Japan's national dental licensing examination, a wide range of histopathological images are used to answer diagnostic questions; these images are considered reliable regarding both image quality and area selection.

This study aimed to evaluate how accurately ChatGPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet — each equipped with image analysis capabilities — can diagnose various oral pathological conditions. By comparing their performance across different diseases and models, this study aimed to clarify the potential of LLMs in the field of oral pathology.

Materials and methods

Study design and dataset

This study analyzed the 100th to 118th editions of the Japanese National Dental Examination. Each examination comprised approximately 360 questions assessing essential knowledge, general dentistry, specialized dentistry, and clinical practice. The essential questions assess the fundamental knowledge and skills necessary for a dentist, whereas the general questions cover basic medicine, epidemiology, and general dentistry. Clinical practice questions focus on examinations, diagnoses, treatments, and procedural sequences in clinical dentistry. These questions included figures or tables, in contrast to the text-only nature of the other question types, and consisted of multiple-choice questions with answer options of 1, 2, 3, 4, or varied. An example of a pathology image-based question is shown in Fig. 1.

This study evaluated the diagnostic capabilities of three LLMs — ChatGPT-4o (ChatGPT, OpenAI, San Francisco, CA, USA), Gemini 1.5 Pro (Gemini, Google, Mountain View, CA, USA), and Claude 3.5 Sonnet (Claude, Anthropic, San Francisco, CA, USA) — using 176 pathology image-based questions from the Japanese National Dental Examination. The image-based evaluation was conducted between March 1 and 30, 2025. Each question included a brief clinical scenario and pathological image. LLMs were prompted to provide a free-text diagnosis. The official diagnosis published by the Ministry of Health, Labour and Welfare was used as the gold standard. The full prompt used for each question was: "Please provide a diagnosis for the oral pathology image, based on the clinical description provided in the National Dental Examination." Each question was submitted in a new session to prevent the influence of prior queries.

Evaluation metrics and statistical analyses

For each model, the proportion of correct responses was calculated for all 176 questions. Cohen's kappa coefficients were used to evaluate the level of agreement between the output of each model and the correct diagnosis. The Landis and Koch scale was applied for the interpretation of

78　37 歳の男性。かかりつけ歯科医を受診した際、エックス線画像で左側上顎の異常像を指摘され来院した。これまでに自覚症状はなく、口腔内外に腫脹も認められない。切除術を行うこととした。初診時のエックス線画像(別冊No. 32 A)、CT(別冊No. 32 B)、切除物の写真(別冊No. 32 C)及び H-E 染色病理組織像(別冊No. 32 D)を別に示す。

診断名はどれか。1 つ選べ。

a　骨形成線維腫

b　線維性異形成症

c　セメント芽細胞腫

d　歯牙腫，集合型〈集合性歯牙腫〉

e　歯牙腫，複雑型〈複雑性歯牙腫〉

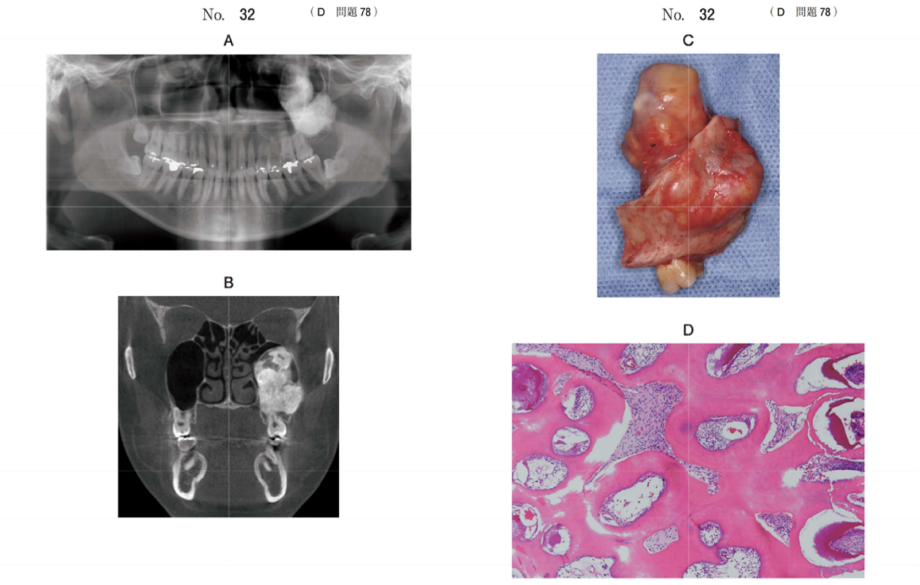

Figure 1　Japanese National Dental Examination (118th Session, Question D78) A 37-year-old man visited our clinic after a primary dentist reported an abnormal radiographic finding of the left maxilla. He had no subjective symptoms and no swelling was observed intraorally or extraorally. Subsequently, surgical excision was performed.
Initial radiographic image (A), CT scan (B), photograph of the excised specimen (C), and H&E-stained histopathological image (D) are shown separately. What is the most likely diagnosis? Choose one. a. Ossifying fibroma b. Fibrous dysplasia c. Cementoblastoma d. Odontoma, compound type e. Odontoma, complex type. CT, computed tomography; H&E, hematoxylin and eosin.

κ-values. McNemar's exact test was used to compare the performance of model pairs (ChatGPT vs. Gemini, ChatGPT vs. Claude, and Gemini vs. Claude) using matched-pair binary outcomes (correct vs. incorrect). To account for multiple comparisons, the Bonferroni correction was considered appropriate and applied to adjust the significance threshold, thereby enhancing the statistical validity of the results.

Confusion matrices were constructed to evaluate the multiclass classification performance of each model, both quantitatively and visually. For all 57 oral pathology categories, the model predictions were compared against the ground truth labels provided in each question, and the number of cases for each true-predicted label combination was aggregated.

All statistical analyses were conducted in Python version 3.11.3 (Python Software Foundation, Wilmington, DE, USA) using the scikit-learn, statsmodels, scipy, matplotlib, and seaborn libraries. Statistical significance was set at $P < 0.05$ and applied throughout.

Results

Comparison of diagnostic accuracy

Among the 57 pathological categories evaluated, considerable variability in diagnostic accuracy was observed across the lesion types and models (Table 1). Besides squamous cell carcinoma (SCC), which all models identified

Table 1 Comparison of large language model (LLM) diagnostic accuracy by oral pathological category.

Pathological category	Total	Number of correct answers			Accuracy rate (%)		
		ChatGPT	Gemini	Claude	ChatGPT	Gemini	Claude
Squamous cell carcinoma (well differentiated)	19	17	19	16	89.47	100	84.21
Ameloblastoma	14	4	7	5	28.57	50	35.71
Odontogenic keratocyst	11	5	7	4	45.45	63.64	36.36
Squamous cell carcinoma	8	5	6	2	62.5	75	25
Pleomorphic adenoma	8	2	7	2	25	87.5	25
Adenoid cystic carcinoma	8	1	2	0	12.5	25	0
Odontogenic myxoma	6	3	2	3	50	33.33	50
Mucoepidermoid carcinoma	5	0	3	1	0	60	20
Malignant lymphoma	5	4	4	5	80	80	100
Dentigerous cyst	4	0	0	0	0	0	0
Calcifying odontogenic cyst	4	0	1	2	0	25	50
Lichen planus	4	4	4	4	100	100	100
Candidiasis	4	4	4	4	100	100	100
Pemphigus	4	1	1	3	25	25	75
Pemphigoid	4	1	2	3	25	50	75
Sequestrum	4	2	1	2	50	25	50
Lymphoepithelial cyst	3	0	1	0	0	33.33	0
Epithelial dysplasia	3	2	0	2	66.67	0	66.67
Mucocele	3	2	3	1	66.67	100	33.33
Warthin tumor	3	2	1	3	66.67	33.33	100
Hyperkeratosis	3	2	1	1	66.67	33.33	33.33
Fibrous dysplasia of bone	3	1	3	1	33.33	100	33.33
Radicular cyst	2	0	1	2	0	50	100
Schwannoma	2	2	1	2	100	50	100
Melanin pigmentation	2	0	1	0	0	50	0
Hemangioma	2	1	2	2	50	100	100
Amyloidosis	2	2	2	2	100	100	100
Candida	2	2	2	2	100	100	100
Papilloma	2	2	1	1	100	50	50
Fibroma	2	0	2	1	0	100	50
Epidermoid cyst	2	0	0	2	0	0	100
Malignant melanoma	2	2	2	1	100	100	50
Dermoid cyst	2	0	1	1	0	50	50
Eruption cyst	1	0	0	0	0	0	0
Postoperative maxillary cyst	1	0	0	0	0	0	0
Osteosarcoma	1	1	1	0	100	100	0
Complex odontoma	1	0	0	0	0	0	0
Osteomyelitis	1	0	1	0	0	100	0
Ossifying fibroma	1	0	1	0	0	100	0
Pigmented nevus	1	0	0	1	0	0	100
Cemento-ossifying Fibroma	1	0	1	1	0	100	100
GVHD	1	1	1	1	100	100	100
Odontoma	1	0	0	1	0	0	100
Adenomatoid odontogenic tumor	1	0	1	0	0	100	0
Lipoma	1	1	1	1	100	100	100
Fibrous dysplasia	1	0	1	0	0	100	0
Tuberculosis	1	0	1	1	0	100	100
Calcified degeneration	1	0	1	0	0	100	0
Synovial osteochondromatosis	1	0	0	1	0	0	100
Orthokeratinized odontogenic cyst	1	1	1	1	100	100	100
Chronic sialadenitis (Sjögren's syndrome)	1	1	1	1	100	100	100
Chronic diffuse sclerosing osteomyelitis	1	1	1	1	100	100	100
Lymphangioma	1	0	0	1	0	0	100
Cemento-osseous Fibroma	1	0	0	1	0	0	100
Cementoblastoma	1	0	0	0	0	0	0

Table 1 (*continued*)

Pathological category	Total	Number of correct answers			Accuracy rate (%)		
		ChatGPT	Gemini	Claude	ChatGPT	Gemini	Claude
Epulis	1	0	0	0	0	0	0
Nasolabial cyst	1	1	1	0	100	100	0
Overall	176	74	100	84	45.5	61.4	52.3

Table 2 Agreement between each large language model (LLM) and the gold standard diagnosis assessed by Cohen's kappa coefficient.

LLM	κ-value
ChatGPT	0.443
Gemini	0.600
Claude	0.513

Table 3 Pairwise comparisons of diagnostic accuracy between large language model (LLM) using McNemar's exact test.

LLM	*P*-value
ChatGPT vs. Gemini	<0.001
ChatGPT vs. Claude	0.119
Gemini vs. Claude	0.072

with high accuracy, several other lesions — including candidiasis, lichen planus, amyloidosis, and chronic inflammatory conditions (e.g., chronic sialadenitis and sclerosing osteomyelitis) — were correctly diagnosed by all three LLMs. By contrast, cystic lesions such as dentigerous cysts, eruption cysts, and postoperative maxillary cysts exhibited consistently low accuracy, with none of the models being able to correctly identify these entities (see Tables 2,3).

Gemini demonstrated superior performance in diagnosing pleomorphic adenomas (87.5 %) and ameloblastomas (50 %), whereas ChatGPT and Claude showed markedly lower accuracies for these tumor types. Adenoid cystic carcinoma proved challenging for all models, particularly Claude, which yielded 0 % accuracy. Claude also showed relatively low accuracy for ameloblastoma and odontogenic keratocysts (35.7 % and 36.4 %, respectively), underscoring the differences in model-specific diagnostic capabilities.

Overall, Gemini outperformed the other models in most categories, particularly in the diagnosis of odontogenic and salivary gland tumors. ChatGPT and Claude demonstrated comparable performance, with slight variations depending on lesion type.

Agreement with the gold standard

Cohen's kappa coefficients were calculated to assess the level of agreement between each model's predictions and gold standard diagnoses. ChatGPT, Gemini, and Claude achieved κ-values of 0.443, 0.600, and 0.513, respectively. According to the Landis and Koch scale, all three models demonstrated moderate agreement, with the Gemini model approaching the threshold for substantial agreement. These results suggest that Gemini exhibited the highest consistency with the reference standard. As many categories had $\leq$4 examples, κ-values should be interpreted with caution due to class imbalance.

Inter-model comparison

Pairwise comparisons between models were performed using McNemar's exact test; Gemini and ChatGPT showed statistically significant differences ($P < 0.001$), favoring Gemini. No significant differences were observed between Claude and ChatGPT ($P = 0.193$), or between Claude and Gemini ($P = 0.064$), although the latter showed a trend toward the superiority of Gemini.

Confusion matrix analysis

To visualize diagnostic errors and identify frequently confusing categories, confusion matrices were constructed for each model. ChatGPT frequently misclassifies odontogenic tumors as cystic lesions. Gemini exhibited fewer overall misclassifications but still showed a notable overlap in glandular tumor diagnoses. Claude tended to confuse soft tissue and salivary gland tumors more frequently. These findings suggest model-specific differences in the interpretation of visual patterns and distinction of closely related entities.

Analysis of the misclassification patterns revealed distinct tendencies among the models. For ChatGPT, the most frequent confusion was between pleomorphic adenoma and adenoid cystic carcinoma (six cases), and in the reverse direction (five cases). Other notable errors included pleomorphic adenoma misclassified as mucoepidermoid carcinoma and lichen planus misclassified as pemphigus (3 cases each; Fig. 2).

In Gemini, the most common misclassification occurred between adenomatoid odontogenic tumors and ameloblastomas (four cases), followed by pleomorphic adenoma and adenoid cystic carcinoma (three cases; Fig. 3). For Claude, five misclassification pairs were observed, each occurring in three cases: Warthin tumor versus adenoid cystic carcinoma, dentigerous cyst versus odontogenic keratocyst, odontogenic keratocyst versus mucoepidermoid carcinoma, adenoid cystic carcinoma versus pleomorphic adenoma, and adenomatoid odontogenic tumor versus ameloblastoma (Fig. 4).

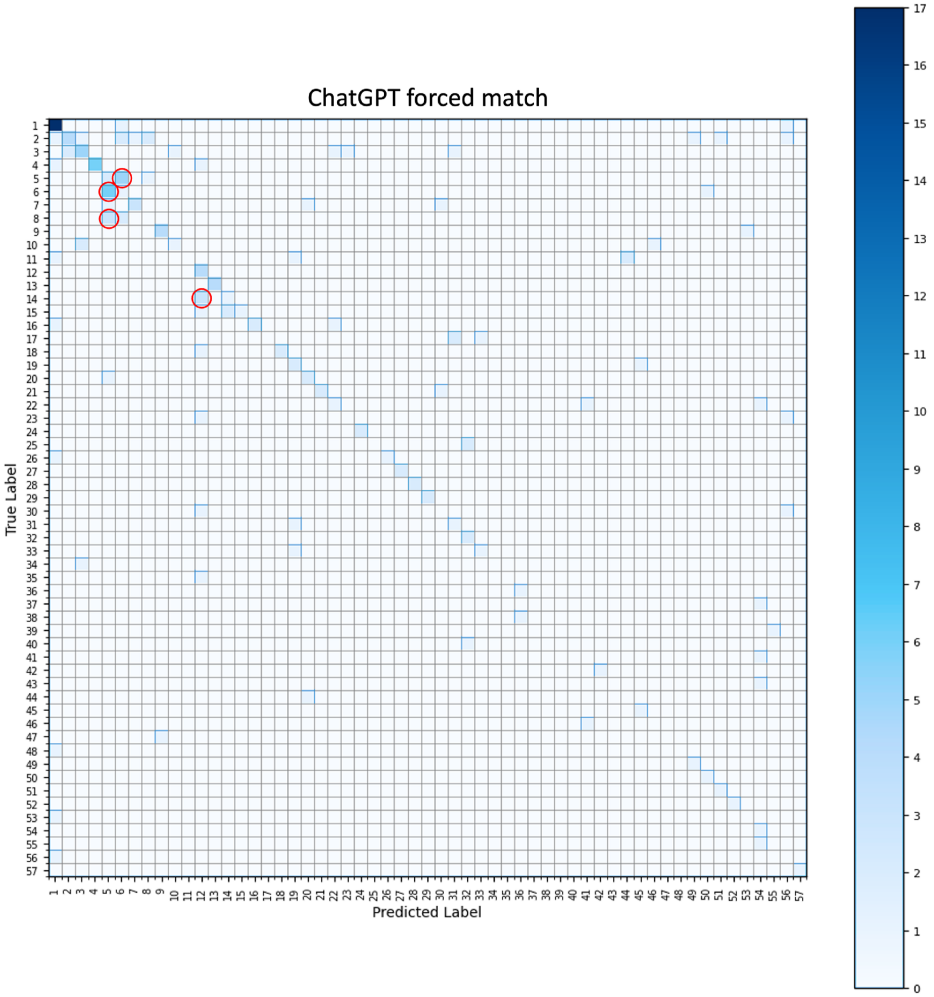

Figure 2 Confusion matrix illustrating the classification performance across 57 oral pathology categories (ChatGPT).
This matrix compares the true labels (vertical axis) with the predicted labels (horizontal axis) for each diagnostic category. The color intensity reflects the number of cases per cell, with darker shades indicating higher frequencies. The red circles indicate the answers that were most frequently answered incorrectly. The labels are as follows: 1. Squamous cell carcinoma (well-differentiated), 2. Ameloblastoma, 3. Odontogenic keratocysts, 4. Squamous cell carcinoma, 5. Pleomorphic adenoma, 6. Adenoid cystic carcinoma, 7. Odontogenic myxoma, 8. Mucoepidermoid carcinoma, 9. Malignant lymphoma, 10. Dentigerous cyst, 11. Calcifying odontogenic cyst, 12. Lichen planus, 13. Candidiasis, 14. Pemphigus, 15. Pemphigoid, 16. Sequestrum, 17. Lymphoepithelial cyst, 18. Epithelial dysplasia, 19. Mucocele, 20. Warthin tumor, 21. Hyperkeratosis, 22. Fibrous dysplasia of bone, 23. Radicular cyst, 24. Schwannoma, 25. Melanin pigmentation 26. Hemangioma, 27. Amyloidosis, 28. Candida, 29. Papilloma, 30. Fibroma, 31. Epidermoid cyst, 32. for malignant Melanoma, and 33. Dermoid cyst, 34. Eruption cyst, 35. Postoperative maxillary cyst, 36. Osteosarcoma, 37. Complex odontoma, 38. Osteomyelitis, 39. Ossifying fibroma, 40. Pigmented nevus, 41. Cemento-ossifying fibroma, 42. GVHD, 43. Odontoma, 44. Adenomatoid odontogenic tumors, 45. Lipoma, 46. Fibrous dysplasia, 47. Tuberculosis, 48. Calcified degeneration 49. Synovial osteochondromatosis, 50. Orthokeratinized odontogenic cysts 51. Chronic sialadenitis (Sjögren's syndrome), 52. Chronic diffuse sclerosing osteomyelitis, 53. Lymphangioma, 54. Cemento-osseous fibromas, 55. Cementoblastoma, 56. Epulis, 57. Nasolabial cyst.

Discussion

In recent years, the application of LLMs in the medical field has attracted considerable attention. However, their practical utility in dentistry, particularly in image-based pathological diagnosis, has not been sufficiently validated. This study evaluated the diagnostic performance and potential educational applications of LLMs by using image-based questions from the Japanese National Dental Examination. This study investigated the diagnostic capabilities of three LLMs (ChatGPT, Gemini, and Claude) using 176 image-based pathology questions from the Japanese National Dental Examination. All the models demonstrated moderate agreement with the gold standard, with Gemini showing the highest overall accuracy and significantly outperforming ChatGPT. This finding underscores the capacity of LLMs to address image-based questions in the National Dental Examination, as well as provides evidence of their potential diagnostic performance in the field of oral pathology. However, it should be noted that the images used in the national examination were selected and standardized by experts and may therefore disproportionately

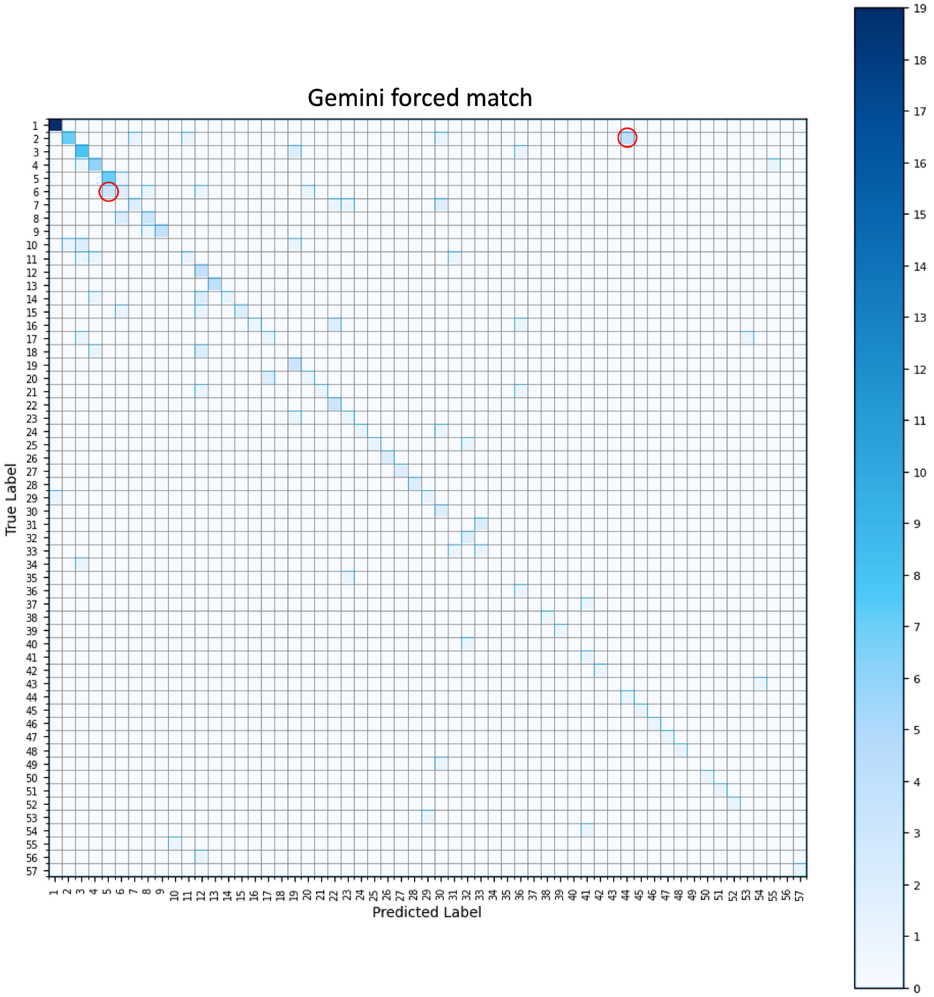

Figure 3 Confusion matrix illustrating the classification performance across 57 oral pathology categories (Gemini). This matrix compares the true labels (vertical axis) with the predicted labels (horizontal axis) for each diagnostic category. The color intensity reflects the number of cases per cell, with darker shades indicating higher frequencies. The red circles indicate the answers that were most frequently answered incorrectly.

represent diagnostically straightforward cases. This introduces a potential limitation in external validity, as real-world clinical images often involve more complex or ambiguous presentations. Nevertheless, even with Gemini — which achieved the highest accuracy of 100 % for SCC — the overall accuracy varied substantially across lesion types, with the lowest being 38.7 % for cystic lesions. According to a recent study, ChatGPT-4o demonstrated the lowest performance in diagnosing granuloma and inflammation cases (100 % incorrect), while achieving the highest accuracy in mucocele cases (93.3 % correct).[11] This variability suggests that further validation is required before these models can be reliably applied in clinical oral pathology.

Previous research assessed LLMs in text-based clinical contexts, primarily focusing on multiple-choice formats. ChatGPT-4o has been shown to outperform ChatGPT-3.5 in the Japanese dental board examination,[12,13] demonstrating high accuracy, particularly in basic science domains such as pathology and pharmacology.[13] However, these studies excluded image-based or multimodal questions. By contrast, the present study uniquely evaluated the performance of LLMs with multimodal inputs, simulating real-world diagnostic scenarios involving clinical images and narratives, an area underexplored in dental AI research.

The application of AI in pathological diagnosis has advanced, primarily in the field of digital pathology, where it plays a supportive role. Utilizing whole-slide imaging, AI systems have been employed to analyze high-resolution images for the quantification of cellular components and immunohistochemical markers, as well as to assess the morphological features of tumor tissues. These applications are expected to contribute to the standardization and reproducibility of diagnoses and reduce diagnostic time. More recently, AI systems categorized as "independent reporting algorithms" have been developed to autonomously generate diagnostic results without the intervention of pathologists. These include algorithms for breast cancer grading and lymph node metastasis detection.[1] To the best of our knowledge, the present study represents the first report to explore the current potential of AI for automated diagnosis of oral lesions.

When analyzing the results by disease category, all LLMs demonstrated a high diagnostic accuracy for SCC, with

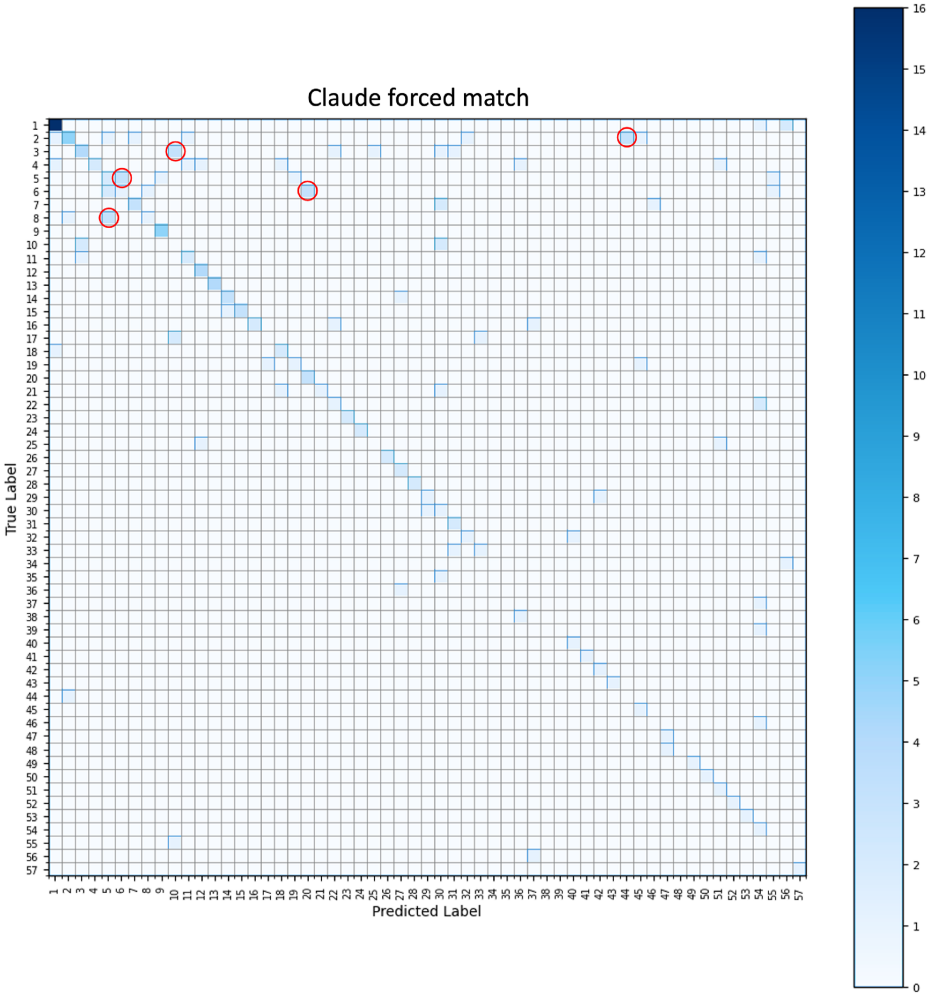

Figure 4 Confusion matrix illustrating the classification performance across 57 oral pathology categories (Claude).
This matrix compares the true labels (vertical axis) with the predicted labels (horizontal axis) for each diagnostic category. The color intensity reflects the number of cases per cell, with darker shades indicating higher frequencies. The red circles indicate the answers that were most frequently answered incorrectly.

scores of 89.5 %, 100 %, and 84.2 % for ChatGPT, Gemini, and Claude, respectively. Notably, Gemini achieved a perfect score (100 %) for well-differentiated SCC. SCC is the most common malignant tumor of the oral cavity[14] and frequently appears in the Japanese National Dental Examination. Thus, these LLMs can be effectively utilized as self-learning aids, particularly for SCC. Moreover, in settings where oral pathology specialists are unavailable, Gemini Advanced could serve as a diagnostic support tool, specifically for this disease. However, in such cases, it is essential to ensure that appropriate biopsy specimens are obtained and that the correct tissue sections are analyzed; therefore, careful consideration and future improvements are warranted. Still, since diagnostic accuracy is not absolute, the ultimate clinical responsibility rests with the attending dentist, as previously discussed.[15]

A confusion matrix analysis of incorrect answers revealed common misclassifications among the three LLMs. It should be noted that Cohen's κ is sensitive to class imbalance. In this dataset, where many classes have $n \leq 4$ correct answers, κ values may be disproportionately affected by class frequencies. Therefore, caution is required when interpreting κ values in this context. One frequent source of error was the differentiation between pleomorphic adenoma and adenoid cystic carcinoma, with ChatGPT identifying 11 cases, the highest among the models, and Gemini identifying three cases (the second highest). Claude also identified three cases, making this one of the most common sources of error across all models. Misclassifications also frequently occurred between pleomorphic adenoma and mucoepidermoid carcinoma, with three cases in ChatGPT and three cases in Claude. These tumors are all salivary gland neoplasms and share certain histological features, such as duct-like structures, which may have contributed to diagnostic confusion.[16] Additionally, cases of ameloblastoma were sometimes misclassified as adenomatoid odontogenic tumors by both Gemini and Claude. Improving the diagnostic accuracy of similar lesions is therefore essential for the future development of reliable LLMs.

The observed differences in diagnostic accuracy among LLMs may stem from several factors: variations in the content and quality of the datasets used to train each model, which can affect their understanding of specific diseases or image patterns; differences in multimodal processing

capabilities, which can influence their ability to integrate and interpret image and textual data; differences in reasoning styles, as each model may adopt distinctive approaches to information interpretation and decision making; differences in response generation strategies, including the prioritization of safety, consistency, and explainability, which can impact the confidence and clarity of the diagnoses; and differences in the ability to distinguish between visually similar diseases, such as salivary gland tumors and cystic lesions, which may lead to inconsistent diagnostic outcomes.[17] To improve accuracy, it is important to train models with diverse, high-quality datasets to enhance their ability to process images and text in an integrated manner, and implement transparent reasoning and expert validation to reduce misclassifications.

These findings highlight the potential of LLMs as supportive tools in dental education and clinical decision making. They can assist students in self-directed learning and support general practitioners in formulating differential diagnoses in low-resource environments. However, the variability in reasoning approaches, risk of hallucinated content, and occasional overconfident diagnostic predictions observed in both the present and prior studies underscore the necessity of human oversight.

Future work should explore standardized evaluation frameworks for multimodal diagnostic performance, transparency-enhancing strategies (such as explainable AI), and dynamic tracking of performance across model updates. Furthermore, the integration of laboratory findings, medical histories, and structured patient data may help mitigate reasoning fallacies and improve diagnostic fidelity in real-world applications.

In conclusion, this study evaluated the diagnostic capabilities of three LLMs using 176 pathology image-based questions from the Japanese National Dental Examination. All models demonstrated moderate agreement with the gold-standard diagnoses; however, Gemini Advanced achieved the highest diagnostic accuracy and statistically significant superiority over ChatGPT.

By incorporating image-based questions, this study provides a clinically realistic assessment of the multimodal diagnostic abilities of LLMs, distinguishing it from prior research that has primarily focused on text-only, multiple-choice formats. The performance differences observed among the models, as well as the variation in accuracy across diagnostic categories, highlight the importance of understanding the strengths and limitations of each model before clinical application.

These findings underscore the potential of LLMs as supportive tools in dental education and clinical decision making. Nonetheless, appropriate safeguards — including human oversight and explainability — are essential to ensure safe and effective integration into healthcare practices. Further research using real-world clinical datasets and more sophisticated evaluation frameworks is vital for advancing AI-supported diagnostics in dentistry.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Acknowledgments

This study was supported by research funding from the Health Sciences University of Hokkaido, Japan.

References

1. Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. *Diagn Pathol* 2023;18:109.
2. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—a recent scoping review. *Diagn Pathol* 2024;19:43.
3. Ding L, Fan L, Shen M, et al. Evaluating ChatGPT's diagnostic potential for pathology images. *Front Med* 2024;11:1507203.
4. Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature* 2024;634:466–73.
5. Cascinelli N, Ferrario M, Tonelli T, Leo E. A possible new tool for clinical diagnosis of melanoma: the computer. *J Am Acad Dermatol* 1987;16:361–7.
6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
7. Kim JS, Kim BG, Hwang SH. Efficacy of artificial intelligence-assisted discrimination of oral cancerous lesions from normal mucosa based on the oral mucosal image: a systematic review and meta-analysis. *Cancers (Basel)* 2022;14:3499.
8. Li J, Kot WY, McGrath CP, Chan BWA, Ho JWK, Zheng LW. Diagnostic accuracy of artificial intelligence assisted clinical imaging in the detection of oral potentially malignant disorders and oral cancer: a systematic review and meta-analysis. *Int J Surg* 2024;110:5034–46.
9. Zhan ZZ, Xiong YT, Wang CY, et al. Utilizing GPT-4 to interpret oral mucosal disease photographs for structured report generation. *Sci Rep* 2025;15:5187.
10. Diniz-Freitas M, Rivas-Mundiña B, García-Iglesias JR, García-Mato E, Diz-Dios P. How ChatGPT performs in oral medicine: the case of oral potentially malignant disorders. *Oral Dis* 2024;30:1912–8.
11. Cuevas-Nunez M, Silberberg VIA, Arregui M, et al. Diagnostic performance of ChatGPT-4o in histopathological description analysis of oral and maxillofacial lesions: a comparative study with pathologists. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2025;139:453–61.
12. Morishita M, Fukuda H, Yamaguchi S, et al. An exploratory assessment of GPT-4o and GPT-4 performance on the Japanese National Dental Examination. *Saudi Dent J* 2024;36:1577–81.
13. Uehara O, Morikawa T, Harada F, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese national dental examination. *J Dent Educ* 2025;89:459–66.
14. Badwelan M, Muaddi H, Ahmed A, Lee KT, Tran SD. Oral squamous cell carcinoma and concomitant primary tumors, what do we know? A review of the literature. *Curr Oncol* 2023;30:3721–34.
15. Mathew MG, Jose S, Yadav S, Cherian J. AI in dentistry: a legal minefield? *Br Dent J* 2025;238:838.
16. Speight PM, Barrett AW. Salivary gland tumours. *Oral Dis* 2002;8:229–40.
17. Sonoda Y, Kurokawa R, Hagiwara A, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. *Jpn J Radiol* 2025;43:586–92.